Chapter 7
Categorical Variables

Section 7.2
Distribution of Data Values for Two Categorical Variables

David W. Gerbing

The School of Business
Portland State University

- Distribution of Data Values for Two Variables
  - Joint Frequencies
  - Probabilities
  - Ordinal Data
  - Chart Direct from Frequencies

7.2a
Joint Frequencies

How Often Do Values of Two Variables Occur Together?

Joint Frequencies

- Some questions of interest to the manager
  - For relation between Supplier and Quality: How many parts from Supplier A are Defective?
  - For relation between Government Pay Grade and Gender: What % of employees at each pay grade are women?
  - To help plan inventory for a road show, what proportions of different style jackets do riders of a particular brand of motorcycle purchase
- To study the relationship between the values of two categorical variables, the key concept is the joint frequency
- **Joint Frequency**: The count of how often the same combination of values occur on each of two variables
- The joint distribution of two categorical variables can be displayed as a table or a graph, as for a single variable

Illustration: Joint Frequencies

Type of Motorcycle and Jacket Thickness

- Consider Type of Motorcycle and Thickness of Jacket
  - Type of Motorcycle, or Bike, operationalized as a categorical variable with two values: Sport, Touring
  - Thickness of Jacket Material operationalized as a categorical variable with three values: Lite, Med, Thick
- How to stock the different Jacket Thicknesses at a vendor booth at a motorcycle show?
- Touring riders are on more stable bikes and are presumed less likely to be concerned with protection from falls
- So there is likely a relationship between these variables

  The preference for Jacket Thickness tends to change across Type of Motorcycle
- The purpose of the analysis is to examine past sales data to see if the preference exists

The Data

Begin with the data table

- The data are organized into a data table stored as, for example, a csv text file or Excel file
- Consider a data table of 443 past sales, with the Type of Motorcycle and Jacket Thickness recorded for each sale
- For this csv data file, bike.csv, the first line contains the variable names, here followed by the first three lines of data

```
Bike,Jacket
Sport,Thick
Touring,Lite
Touring,Thick
...
```

- Read this data table into R, as the data frame `d`

```
> d <-
  Read("http://web.pdx.edu/~gerbing/data/bike.csv")
```

Joint Frequencies

From data to the table of counts

- The statistical analysis of counting the joint occurrences of the values of the two variables results in a *table* of counts
- Another name for counting is tabulation
- **Cross-Tabulation Table** or **Cross-Tab Table**: Table of the joint frequencies of the values of two or more categorical variables
- For two categorical variables, can refer to the cross-tabulation table as a two-way cross-tabulation table

Joint Frequencies from lessR

From data to the table of counts

- Is there a relation between the two variables Style of Motorcycle Jacket and Type of Motorcycle?

- R: Each of the following statements yields the following joint frequency distribution
  > `BarChart(Bike, by=Jacket)`, or abbreviate with `bc()`
  > `Plot(Bike, Jacket)`, or abbreviate with `sp()`
  > `SummaryStats()`, or `ss()`, no graph, but different versions of the joint frequency table

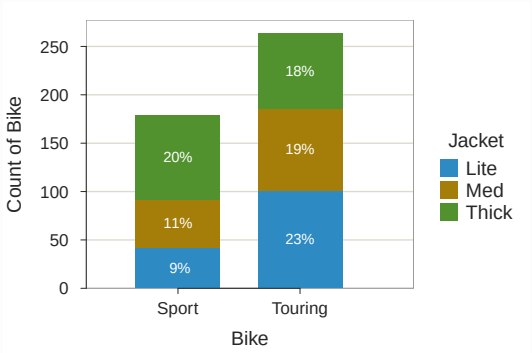
```
        Bike
Jacket  Sport Touring
  Lite     42     101
  Med      50      85
  Thick    87      78
```

e.g., the joint frequency of 42 is how many Sport motorcyclists chose a Lite jacket

Stacked Bar Chart of Two Categorical Variables

- Stacked bar chart, focus on the size of each group for the x-axis variable, bars have the same shape as for a single variable

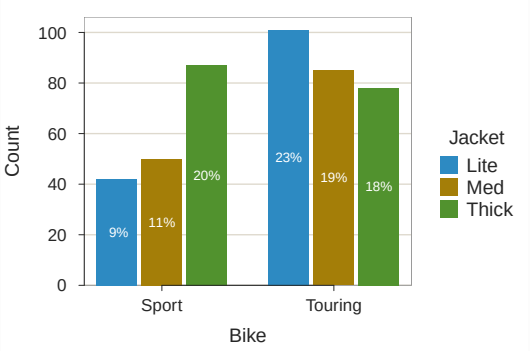
```
> BarChart(Bike, by=Jacket)
```



Grouped Bar Chart of Two Categorical Variables

- Grouped bar chart, focus on comparing levels to each other with the beside parameter

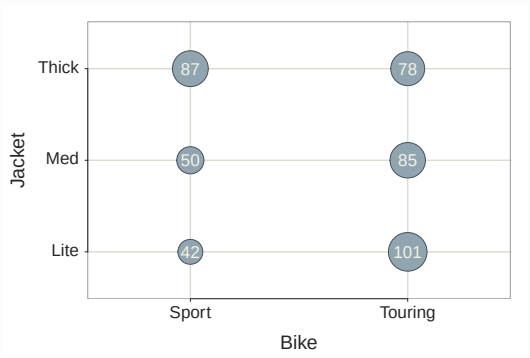
```
> BarChart(Bike, by=Jacket, beside=TRUE)
```



Scatterplot of Two Categorical Variables

- Bubble chart, the size of a bubble corresponds to the frequency

```
> Plot(Bike, Jacket)
```



Description and Inference

Statistical analysis

- These functions also provide analysis in addition to graphics
- Describe the extent of the relationship
  - Cramer's V is a kind of correlation coefficient, which varies from 0 for no relationship to 1 for perfect relationship

```
Cramer's V: 0.203
```

  - A moderate relationship detected in this sample

- Inference is with the chi-square test
  - Null Hypothesis: No relationship

```
Chi-square Test:
    Chisq = 18.271, df = 2, p-value = 0.000
```

  - Reject the null because p-value < 0.05
  - Conclude: Jacket Thickness and Type of Motorcycle related

Illustration: Managerial Conclusion

Identify the relationship between Motorcycle and Jacket

- Two variables are related if information regarding the value of one variable allows for a better prediction as to the value of the other variable
- At least as a description of the current data, Type of Motorcycle appears related to Jacket Thickness
- Examine the results . . .
  - For riders of Sport Motorcycles, there is an increasing trend for increased Jacket Thickness and a similar decreasing trend, for Touring riders, though not quite as pronounced
  - For Sport riders, the counts and column %'s increase for increasing jacket thickness from 42 (23.5%) to 50 (27.9%) to 87 (48.6%), while for Touring riders, the decrease is from 101 (38.3%) to 85 (32.2%) to 78 (29.5%)
- Note that these results only describe these data

7.2b
Probabilities

The Margins

Also analyze the separate distribution of each variable

- `BarChart()` and `SummaryStats()` also provide the separate frequency distribution of each variable
- **Marginal Frequency**: A row or column sum from the table of joint frequencies

	Bike		
Jacket	Sport	Touring	Sum
Lite	42	101	143
Med	50	85	135
Thick	87	78	165
Sum	179	264	443

- For example, a total of 143 of all riders, Sport and Touring motorcyclists, chose Lite
- **Grand Total**: Total sample size, in this example, 443

Probabilities from the Joint Frequency Table

Move from counts to proportions

- **Sample Probability**: A proportion, the ratio of observed frequency to total frequency
- One type of probability in this context is the cell probability
- **Cell probability**: A joint frequency divided by the total number of observations, which in this example is 443

	Bike		
Jacket	Sport	Touring	Sum
Lite	0.095	0.228	0.323
Med	0.113	0.192	0.305
Thick	0.196	0.176	0.372
Sum	0.404	0.596	1.000

- e.g, 9.5% of all riders ride a Sport bike wearing a Lite jacket
- **Marginal Probability**: A marginal frequency divided by the corresponding row or column total

Probabilities within Each Column

Proportions within each column or row

- **Conditional Probability**: A proportion from a joint frequency calculated from the corresponding column or row total
- Of interest are the column proportions, the proportion of different Jackets separately sold for each group of bikers
- With this information the vendor can better plan for inventory when selling primarily to primarily Sport or Touring motorcyclists

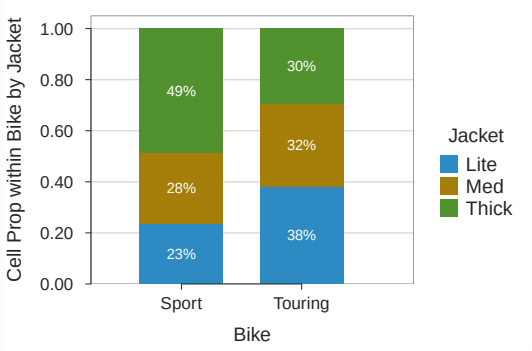
	Bike	
Jacket	Sport	Touring
Lite	0.235	0.383
Med	0.279	0.322
Thick	0.486	0.295
Sum	1.000	1.000

Ex: IF the rider is a Sport biker, 23.5% of the chosen Jackets are Lite, almost half are Thick

100% Stacked Bar Chart

- These conditional probabilities, the column proportions, can also be displayed as a chart from `BarChart()`

```
> BarChart(Bike, by=Jacket, stack100=TRUE)
```



7.2c
Ordinal Data

R: Designate Ordinal Data with an Ordered Factor

Jacket Thickness a progression from Lite to Med to Thick

- Two issues regarding this ordered progression of thickness
  - By default, R presents the values alphabetically in the output
  - It is just a coincidence that the desired order is alphabetical, with the values starting with L, M and T
  - Formally define Jacket Thickness as an ordered factor, in which `Lite<Med<Thick`

- Use the lessR `factors()` function, built with the R `factor()` function, to address these issues
- Replace the variable Jacket with its updated, ordered version
  - To specify order of presentation, invoke the `levels` option
  - To specify ordinal data, also invoke the `ordered` option

```
> d <- factors(Jacket,
         levels=c("Lite","Med","Thick"),
         ordered=TRUE)
```


Bar Chart of Ordinal Data

- lessR `BarChart()`, or `bc()`, plots the frequency bars of ordinal data as an ordered progression of lightness/darkness
- For the ordinal coloring in shades of the same hue to reflect the ordering, the x-axis variable plots the ordinal variable

```
> BarChart(Jacket, by=Bike)
```

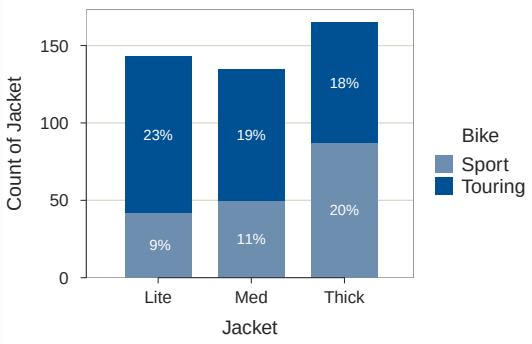


Chart Direct from Frequencies

Bar Chart Directly from the Joint Frequencies I

Enter counts directly as a data frame

- Sometimes the joint frequencies have already been calculated, with the original data not available
- In this situation, to obtain the bar chart, enter the joint frequencies directly, one row per frequency
- A bar chart plots at least one categorical variable, x, against a numerical variable, y, which can represent any numerical entity, not just counts
- Read the numeric y variable directly instead of having `BarChart()` compute its value as frequencies
- Ex: read this data table into R as the data frame `d`

```
> d <- Read("http:
  //web.pdx.edu/~gerbing/data/BikeJacketFreq.xlsx")
```


Bar Chart Directly from the Joint Frequencies II

Structure of the data

- Provide the names of the levels, the row and column names

```
> d
     Bike Jacket Freq
1   Sport   Lite   42
2 Touring   Lite  101
3   Sport    Med   50
4 Touring    Med   85
5   Sport  Thick   87
6 Touring  Thick   78
```

- Obtain the same bar chart as before of Bike on the x-axis with the entered counts on the y-axis, and Jacket the by variable

```
> BarChart(Bike, Freq, by=Jacket)
```

- Numeric y can be any set of numbers, even with decimal digits

Index Subtract 2 from each listed value to get the Slide

- The End